Anchor Environmental, L.L.C.
1716 Government Street, Suite F
Ocean Springs, MS 39564
Phone: (228) 818-9626
Fax: (228) 818-9631

Memorandum

To: Stormwater Technical Team

From: Jason Kase and Carl Stivers

CC:

Date: September 2, 2008

Re: Proposed method for calculating basin-weighted statistics for stormwater data

The objective of this memo is to propose an approach to calculate basin-weighted statistics for stormwater data using the ProUCL statistical package. These statistics (as well as other statistics) will be used in estimation of stormwater loads, which will be presented in the RI fate and transport loading analysis including in-river abiotic fate and transport modeling. At the last Stormwater Technical Team conference call, EPA tentatively agreed to calculating basin-weighted statistics for stormwater chemical concentrations by land use type provided that 1) other unweighted statistics were also calculated and 2) EPA could review and approve details of such a basin weighting approach. This memo fulfils this second requirement.

Unweighted Data Methods Summary

Before describing methods involving weighted data, it is worth summarizing the un-weighted methods that have also been discussed by the Stormwater Technical Team. In general, the Team has agreed that, to the greatest extent possible, statistical calculations and summarization of the data set will use ProUCL for all data reduction steps. Thus, unweighted data will be input directly into ProUCL, and following the software guidance, summary statistics on each land use's data set will be generated. Per previous meeting agreements, the process for unweighted data analysis can be summarized as:

- Enter unweighted data for each land use into ProUCL 4.0, including non-detects. For data sets with non-detects, ProUCL can create, as appropriate, additional columns to store extrapolated values for non-detects obtained using regression on order statistics (ROS).
- Use ProUCL to conduct goodness-of-fit (GOF) tests to determine distribution of unweighted data.

Commented [KK1]: No. This is the substitution method, which should not be used. Using non-parametric statistics (e.g., ROS), ProUCL can calculate the statistics, but you won't see the values because there are several values that were used for the nondetects to determine the statistics.

Commented [KK2]: While I agree that this step should be conducted, the data may not fit a distribution and may just be "skewed"; thus, the non-parametric methods are best because they do not assume a distribution.

- Use ProUCL graphical displays to present histograms, Q-Q plots, and box plots of unweighted data.
- Use ProUCL to present Summary Statistics and Estimates of Population Parameters for data each land use set consistent with ProUCL recommended methods.

As discussed at the last Stormwater Technical Team meeting, a range of statistics that may be useful in future loading analyses will be generated. This would likely include central tendencies such as means, as well as the 95th percent UCL, and UPL type values.

Commented [KK3]: Upper percentage level?

Weighted Data Considerations

Commented [KK4]: It should be clear that the purpose of this is that some of the stormwater technical team believes that the concentrations are a function of land area runoff within a land use. In order to go forward with this, the first step is to conduct correlation between the parameters (e.g., concentration and flow or concentration and land area – which I did for a limited number of contaminants and found no correlation). Since this process will be conducted for all contaminants, the correlation should hold true for all contaminants; however, may be that certain classes of chemicals may correlate and then the weighting would be appropriate for only those classes of chemicals.

It should be noted that ProUCL is not intended for estimating statistics on weighted data. Adapting ProUCL to these purposes may have unintended biases or outcomes in the data analysis that are difficult to identify. Consideration should also be given to simpler approaches for handling weighted data outside ProUCL, such as calculating geometric means using half detection limits. (In this case, ProUCL would still be used for summarization of unweighted data.) Although these methods may lack technical sophistication, this may be balanced by a clearer understanding of their inherent biases and limitations.

Commented [KK5]: EPA does not support this approach. The geometric mean on skewed data will give the median, which is lower than the mean. Further, EPA does not support using half the detection limits as a substitution for censored data. If the LWG is looking to broaden the range of data using simpler methods, I would suggest using percentiles. This could only be done outside ProUCL with non-detect data since it is inappropriate to use the substitution method to estimate censored data.

Commented [KK6]: It is unclear how this method would allow for understanding inherent biases and limitations. This statement should either be supported or stricken.

In general, two different approaches for basin-weighted statistics have been discussed during previous Stormwater Technical Team conference calls. They are:

1. Calculate summary statistics (e.g., mean, UCL, UTL) using unweighted data and then weight the statistic of interest.
2. Weight each record and use the weighted data to calculate summary statistics.

Commented [KK7]: This should not be done and I don't remember this being discussed.

The first approach has one primary disadvantage in that record-, station-, location-, and basin-specific weighting are not possible. Only a single 'weighting' value may be used on the statistic of interest. Possibly, summary statistics could be generated on a smaller scale (e.g. station-specific) and then weighted on a station-specific basis. However, the number of observations is generally not sufficient to permit such an approach. A minimum of 8-10 observations (more if the frequency of detection is low; < 50%) are generally required before measures of central tendency (i.e., UCL95) can be calculated with a reasonable level of confidence. Thus, approach #1 is not a preferred method.

Commented [KK8]: This approach was discussed, but EPA does not support this because it would limit the number of data points (5 vs. 15) and restrict the use of statistics (statistics really should not be done on less than 10 data points). This is why the premise for the data collection was that each data point would be representative of the land use; otherwise, we would need a minimum of 10 data points from 10 outfalls for each land use.

Commented [KK9]: Approach number 1 cannot be done unless additional data is collected.

Approach #2 has the advantage of allowing each record to be weighted individually based upon more discrete site-specific considerations (e.g., basin size). The primary disadvantage of this approach is that, for data sets with below detection limit (i.e., non-detect) data, some method of representing these data in weighted form must be devised.

Proposed Weighted Data Methods

To allow for more discrete weighting of stormwater data, approach #2 above is preferred. The primary issue to address for this approach is non-detect handling.

Data Sets without Non-Detect Data

For data sets without non-detect data, the proposed method is simple:

1. Apply weights to records (as described below)
2. Test data distributions for weighted and unweighted data sets (for comparison purposes)
3. Calculate summary statistics as desired (e.g., mean, UCL95, etc.) of the weighted data set using the recommended method per ProUCL guidance.

Testing the distributions and summary statistics of weighted and unweighted data (Step 2) would allow for a qualitative assessment of the degree to which the weighting process may have affected the distribution of the data. This step is not considered necessary but may be informative.

Several approaches to weighting data have been discussed by the Stormwater Technical Team. Using the total volume of runoff calculated from an assumed 1-inch amount of precipitation (i.e., equivalent to the unit hydrograph) for each basin sampled would be a simple method of weighting each station's samples relative to flow expected from each basin. Because the total runoff volume will be proportional to the Runoff Coefficient (RC) times Basin Area (A) for each basin, under this approach it does not matter how much rainfall is assumed (i.e., the weighting will not change if for example, 1 inch vs. 2 inches of rain are assumed). The proposed weighting method will also correct for differing numbers of samples that may exist for each basin by using the average concentration from each sampled location. In summary, to obtain a weighted concentration (Cw), calculate per the following steps:

1. The average concentration (C) for each sample location

Commented [KK10]: The weighting factor can be pre-established and multiplied by each value (detect or nondetect) and input into ProUCL where nondetects are flagged and allow ProUCL to conduct the statistic. The pre-weighting could use a method such as summing the volume discharged within a land use for each sample collected and then determining the fraction for each sample by dividing the volume discharged for the sample by the total volume discharged. Again, there should be shown that the parameter that is going to be used for weighting should have correlation with the concentration.

Commented [KK11]: The rational for this needs to be more than just that it is preferred because it is undesirable to collect more data. It is more appropriate to describe the conceptual model and the method that conforms to that conceptual model. If the conceptual model is that the larger basins are more representative of central tendency of the land use type throughout the site, then the appropriate approach is to collect more information and determine the variability in each basin in extrapolating to other sites. Alternatively, if the conceptual model is that each basin is equally representative of the land use, then the appropriate approach would be to use the existing data set. Weighting of data should only be done when it is shown that there is a correlation between two or more parameters (e.g., concentration and flow, or concentration and flow and land area, or concentration and flow and impervious area, etc.).

Commented [KK12]: The only way to do this is to pre-determine the weighting as discussed above.

Commented [KK13]: This method can be used for censored data sets, as well.

Commented [KK14]: The summary statistics should be compared, as well. Summary statistics will provide more useful information than the distributions.

Commented [KK15]: EPA disagrees with this statement. This step provides critical information regarding the effects of the weighting of the data.

Commented [KK16]: These should be discussed if you are going to make this statement.

Commented [KK17]: Is this the approach that the LWG is proposing? If so, they first need to show that there is correlation between this parameter and concentration.

Commented [KK18]: If you pre-weight the data as described above, then this would be accounted for.

Commented [KK19]: This is method 1, above, that the paper stated was not the preferred method. If this method is going to be done, then more data needs to be collected.

2. A weighting factor (W) (unitless fraction) for each sample location based on its runoff volume divided by the sum of volumes for all locations
3. Multiply (C) for each location by W and the number of sample locations (N) in the land use category to obtain Cw (i.e., Cw = C x W x N).

Data Sets with Non-Detect Data

The ROS and KM methods are available in ProUCL for calculating the mean and standard deviation of left-censored datasets (non-detect observations). Following the steps detailed in ProUCL and related guidance, the KM method is always recommended for data sets with non-detects.

To use the KM method with weighted data, several factors will be considered. KM is a nonparametric method that does not require that the data fit a known distribution. In addition, a substitution method for non-detect observations is not required and allow for multiple detection limits within the data set. The details of this approach are not described here but are available in (USEPA 2007, Singh et al. 2006, Bechtel Jacobs 2000). In practice, the mean and standard deviation (as well as UCL95) will be influenced not by the absolute magnitude of non-detect observations but by their frequency and location within the rank order observations (e.g., highest to lowest). Thus, the use of weights may affect the rank order of the observations; influencing the estimate of the mean and standard deviation. The influence of changes on the rank order of detect and non-detect observations is uncertain and would be a function of the individual data sets considered. It is unlikely to have a significant effect on the estimate of the mean unless the probability of detection is strongly correlated with the weights applied (e.g., low concentrations are associated with low weighting). Given that there is not a compelling reason to believe that low concentrations are correlated with low weighting, the KM method to estimate the mean and standard deviation will be applied to weighted data that includes non-detect observations.

To provide a sense of the variability in the estimates of the mean and standard deviation of weighted data sets with non-detect observations using the KM method, parametric estimates of these parameters will be calculated for comparison to the KM estimates. Goodness of Fit (GOF) and graphical methods will be used to best fit the weighted distribution using detected values only. For this comparison, ROS will be the substitution method for parametric estimates of

Commented [KK20]: This will have more effect if there are nondetects that are greater than detects. This will likely have more effect on the sediment trap data than the water data due to increase in detect level resulting from the lack of sample volume.

Commented [KK21]: It may be appropriate that low concentrations are correlated with low weighting – Again, a correlation of the weighting parameter with the concentration needs to be done to show that this is an appropriate weighting factor.

Commented [KK22]: Why only for detected values?

Commented [KK23]: ROS is not a substitution method and is used for censored data. This does not make sense with the previous sentence. Are you saying here that the statistics are going to be used to determine the effects of weighting?

statistics. Although statistics generated using the parametric ROS methods in ProUCL will not be recommended for final loading estimates, it does provide a check on the non-parametric estimates. ROS is recommended for this comparison because it interpolates the non-detects and assigns them a value. Thus, the data set created by this substitution no longer contains non-detects and the potential bias due to changes in rank order of non-detects due to weighting does not exist using ROS. Consequently, the ROS method can be freely used to calculate weighted statistics of interest (e.g., UCL95) by first assigning non-detect values to the raw (unweighted) data and then assigning weights per the above methods. If it appears that the KM method has significant limitations based on this comparative analysis, this will be discussed in the loading estimate results presentation (in the draft RI). Any necessary cautions about use of KM generated statistics for loading estimates will be presented.

Once the mean and standard deviation are estimated using the KM method, the bootstrapping procedures recommended and incorporated into ProUCL will be used to estimate upper confidence limits of the mean or upper probability limits, as desired.

Thus, the method for censored data sets ~~with non-detect observations~~ are:

1. Apply weights to observed values, per methods described above using the detection limits for non-detect data. Identify weighted non-detect values as non-detects per normal data entry into ProUCL.
2. Calculate the mean and standard deviation using KM methods as recommended by ProUCL
3. Calculate the desired statistics (e.g., UCL) using ProUCL (likely performed at same time as Step #2)
4. Use GOF (Shapiro-Wilk) and graphical methods to estimate the distribution of detected data in the dataset
5. Calculate the mean, standard deviation, and desired statistics (e.g., UCL) using the ProUCL recommended methods for that distribution using ROS substitution methods for comparative purposes to KM method
6. Conduct qualitative comparison of statistics calculated using KM method to those using ROS method to support discussion of potential impacts of weighted non-detects on KM method.

Commented [KK24]: I'm unclear what is being said here. The purpose of the parametric ROS method is to get a concentration that will be used to develop the loading estimates for the hybrid model. The parametric ROS uses MLE, yet the document discusses KM.

Commented [KK25]: There will not be a data set created using ROS and this is not a substitution method.

Commented [KK26]: EPA does not support this method – it is proposing a substitution method that is inappropriate for evaluating censored data sets.

Commented [KK27]: I disagree with this statement. KM and ROS are equally valid methods and both preferred to the substitution method. However, you need greater than 50 observations to use KM and should use ROS since there are fewer than 50 observations in the data set.

Commented [KK28]: KM should only be used when you have greater than 50 samples.

Commented [KK29]: KM will not guess below the lowest DL, so if only 1 DL, then KM becomes a substitution method as the DL, which may overestimate the mean. KM will work OK with multiple DL, but should only be used with data sets that have greater than 50 observations.

References

Singh, A., R. Maichle, S.E. Lee. 2006. On the computation of a 95% Upper Confidence Limit of the unknown population mean based upon data sets with below detection limit observations. EPA/600/R-06/022. Office of Research and Development. Washington, DC 20460.

Bechtel Jacobs Company, LLC. Improved methods for calculating concentrations used in exposure assessments. BJC/OR-416. Prepared for the U.S. Department of Energy. Office of Environmental Management.

USEPA. 2007. ProUCL Version 4.0 Technical Guide. EPA/600/R-07/041 Prepared by A. Singh and A. Singh. Office of Research and Development. Washington, DC.